Interspeech 2022
18-22 September 2022, Incheon, Korea

Speech Emotion Recognition via Generation using an Attention-based Variational Recurrent Neural Network

Murchana Baruah, Bonny Banerjee

Institute for Intelligent Systems, and Department of Electrical & Computer Engineering
University of Memphis, Memphis, TN 38152, USA
{mbaruah, bbnerjee}@memphis.edu

Abstract

The last decade has seen an exponential rise in the number of attention-based models for speech emotion recognition (SER). Most of these models use a spectrogram as the input speech representation and the CNN or RNN or convolutional RNN as the key machine learning (ML) component, and learn feature weights to implement attention. We propose an attention-based model for SER that uses MFCC as the input speech representation and a variational RNN (VRNN) as the key ML component. Since the MFCC is of lower dimension than a spectrogram, the model is size- and data-efficient. The VRNN has been used for problems in vision but rarely for SER. Our model is predictive in nature. At each instant, it infers the emotion class and generates the next observation, computes the generation error, and selectively samples (attends to) the locations of high error. Thus, attention emerges in our model, and does not require learning feature weights. This simple model provides interesting insights when evaluated for SER on benchmark datasets. The model can operate on variable length and infinite duration audio files. This work is the first to explore simultaneous generation and recognition for SER, where the generation capability is necessary for efficient recognition.

Index Terms: Speech emotion recognition, recognition by generation, variational RNN, MFCC, attention, active inference, predictive coding.

1. Introduction

Manifestations of emotions are often involuntary. They convey one's true feelings, confidence, intentions, and expectations, which are useful for social interaction. Speech contains linguistic and paralinguistic content; both allow the conveyance of emotions albeit in different ways. The recognition of emotion from speech or *speech emotion recognition* (SER) has been an active area of research in machine learning (ML) for decades, contributing to technologies for human–machine interaction, intelligent tutoring, healthcare, and security.

This paper is concerned with SER without explicitly processing or analyzing the linguistic content in the speech. A large number of ML models have been proposed for this problem (see [1–3] for reviews). Models incorporating attention mechanism often yield superior performance, though at the expense of additional parameters. Attentional models for SER can be largely categorized based on three aspects: speech representation used as input to the ML model, the key ML component in the model, and the implementation of attention.

The input speech is often represented as a linear- or mel-frequency spectrogram (e.g., [4–17]). A few studies have experimented with mel-frequency cepstral coefficients (MFCCs) (e.g., [18]), low-level descriptors (e.g., [19]), or raw speech (e.g., [20]) as input. In [21], the authors investigated the impact of four types of features on the performance of an attentional convolutional neural network (CNN), namely (1) 26 log mel filter banks (logMel), (2) 13 MFCCs, (3) 25 low-level descriptors (frequency- and energy-related parameters and spectral parameters) constituting the extended Geneva minimalistic acoustic parameter set (eGeMAPS), and (4) a prosody feature set consisting of PCM loudness, F0 contour, envelope of F0 contour, voicing probability, local jitter, differential jitter, and local shimmer. From their experiments with the IEMOCAP database, it was concluded that the model and training data (improvised vs. scripted speech) are more impactful than the features.

Models for SER often involve the CNN (e.g., [5, 13, 14, 16–18, 21]), recurrent neural network (RNN) or long short-term memory (LSTM) (e.g., [4, 7–12, 15, 19]), or convolutional RNN/LSTM (e.g., [6, 20]). The variational autoencoder (VAE) has been rarely used (e.g., [22]). Most models utilize *feature-level attention* where the output of hidden layers are weighted. *Decision-level attention* applies to multiple-instance learning [23] where the prediction of the instances are weighted to obtain the bag-level prediction. In both types of attention, the weights are learned along with other parameters to optimize an objective. A CNN with feature-level attention significantly outperforms its decision-level counterpart on benchmark datasets [14].

Contributions. We propose an attention-based model for SER that involves the MFCC and a VRNN. The MFCC is a compressed representation of a lower dimension than a spectrogram, which allows the model to be size- and data-efficient. The unique properties of the proposed model are as follows:
(1) At each instant, the model simultaneously infers the emotion class and generates the observation (input MFCC vector corresponding to the speech window) at the next instant. Training a model by minimizing generation and classification errors in conjunction leads to more stability and less overfitting due to each error acting as a regularizer for the other [24].
(2) Attention emerges in our model due to prediction error. The model selectively samples locations in the input MFCC vector that contain unexpected observations. This saves the use of additional parameters for attention.
(3) Our SER experiments using the RAVDESS and IEMOCAP benchmark datasets show that the proposed end-to-end model yields high classification accuracy by sampling a fraction of the MFCC vector for each window, in addition to providing insights into the model design.

2. Models and Methods

2.1. Preliminaries

Generative model. Given a set of data points x, a generative model p_{model} with parameters θ maximizes the log-likelihood,

10.21437/Interspeech.2022-753

$\mathcal{L}(x;\theta)$, of the data.

Evidence lower bound (ELBO). Let the data x be generated by a latent continuous random variable z. Then, computing the log-likelihood requires integrating the marginal likelihood, $\int p_{model}(x, z)dz$, which is intractable [25]. In variational inference, an approximation of the intractable posterior is optimized by defining an evidence lower bound (ELBO) on the log-likelihood, $\mathcal{L}(x;\theta) \leq \log p_{model}(x;\theta)$.

Variational autoencoder (VAE) is a multilayered generative model. It assumes an isotropic Gaussian prior, $p_\theta(z)$, and i.i.d. data samples. VAE maximizes the following ELBO [25]:

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}[q_\phi(z|x), p_\theta(z)] \tag{1}$$

where $p_\theta(x|z)$ and $q_\phi(z|x)$ are generative and recognition models respectively, $\mathbb{E}$ denotes expectation, and D_{KL} denotes Kullback-Leibler divergence. The first and second terms capture accuracy and complexity respectively. The negative of this ELBO is also known as *variational free energy*, minimization of which has been hypothesized as a general principle guiding brain function [26].

2.2. Problem Statement

Let $\mathbf{X} = \langle X_1, X_2, \ldots, X_T\rangle$ be a sequence of observable variables representing an environment, where T is the sequence length. Let $\mathbf{x}_{\leq t} = \langle x_1, \ldots, x_t\rangle$ $(1 \leq t \leq T)$ be a partial observation of $\mathbf{X}$. Let $\mathbf{y} = \langle y_1, \ldots, y_T\rangle$, where y_t represents the true class label at time t. We define the *prediction* problem as generating $\mathbf{X}$ and $\mathbf{y}$ as accurately as possible from the partial observation $\mathbf{x}_{\leq t}$. At any time t, the objective is to maximize the joint likelihood of X_{t+1} and y_t, given $\mathbf{x}_{\leq t}$ and a generative model p_θ with parameters θ, i.e., $\arg\max_\theta p_\theta(X_{t+1}, y_t|\mathbf{x}_{\leq t})$.

2.3. Proposed Model

Our model comprises of four functions: environment, prediction, action selection, and learning. See Fig. 1.

1. Environment. The environment is the source of sensory data or observations. It is time-varying, consists of variable-length utterances, and can be of infinite duration. Our model interacts with the environment by selectively sampling observations at each time instant.

2. Prediction. The model predicts using a VRNN involving two processes: recognition and generation.

Recognition (Encoder). The probabilistic encoder, $q_\phi(z_t|\mathbf{x}_{\leq t})$, produces a Gaussian distribution over the possible values of the code z_t from which the given observations could have been generated. An RNN with one layer of LSTM units constitute the recognition model. The RNN generates the parameters for the approximate posterior distribution, $q_\phi(z_t|\mathbf{x}_{\leq t})$, from the observations. The prior is sampled from a standard normal distribution, $p_\theta(z_t) \sim \mathcal{N}(0, 1)$, as in [27].

The function of the encoder is shown in Lines 1–3 in Algorithm 2, where RNN_ϕ^{enc} represents the function of an LSTM unit, φ^{enc} is a function that returns the mean and the logarithm of the standard deviation as a linear function of the hidden state, as in [28].

Generation (Decoder). The decoder, $p_\theta(X_{t+1}, y_t|\mathbf{x}_{\leq t}, z_{\leq t})$, generates the perceptual data and the class label from the latent variables, z_t, at each time step. The generative model has two RNNs, each with one layer of hidden LSTM units.

Each RNN generates the parameters of its data distribution. The data is then sampled from this distribution. In our model,

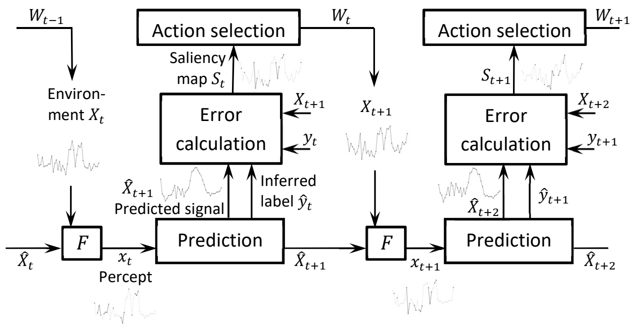


Figure 1: *Flow diagram of the proposed model.*

X_{t+1} is sampled from a multivariate Gaussian with mean and variance generated by the corresponding decoder RNN, and y_t is sampled from a multivariate Bernoulli distribution with mean generated by the corresponding decoder RNN.

The decoder equations are shown in Lines 4–8 of Algorithm 2, where functions RNN_θ^{dec} and φ^{dec} are the same as RNN_ϕ^{enc} and φ^{enc} respectively.

3. Action selection. In the proposed model, action selection amounts to deciding the binary weight (attention) given to each location of the current observation. At any time t, a saliency map S_t is computed from which the action is determined. The saliency map assigns a salience score $S_{t,l}$ to each element l of the input MFCC vector.

The weights are determined by thresholding the prediction error. The threshold is statistically estimated on the fly and is not predetermined.

$$S_t = |X_{t+1} - \hat{X}_{t+1}| \tag{2}$$

$$W_{t,l} = \begin{cases} 1, & \text{if } S_{t,l} \geq \frac{1}{n}\sum_{k=1}^{n} S_{t,k} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$x_{t+1} = W_t \odot X_{t+1} + (\mathbf{1} - W_t) \odot \hat{X}_{t+1} \tag{4}$$

where X_{t+1}, $\hat{X}_{t+1}$ are the true and predicted data (MFCC vectors) respectively, $|.|$ denotes the absolute value, $\odot$ denotes elementwise product, and n is the dimension of a MFCC vector. Eq. 2–4 are written in compact form in Line 7 of Algorithm 1.

Due to the nature of the chosen threshold function, at least one element of the MFCC vector will be salient and at least one element will be non-salient at any time. Our experiments show that variable number of salient features at each time step is more effective. Fixing the number of salient features to a constant occasionally leads to selection of features with low saliency or overlooking features with high saliency. In the proposed model, only the salient MFCC features are sampled. For the non-salient features, the observation at time $t+1$ is the predicted observation from t.

4. Learning. The objective is to maximize the expression in Eq. 5, which can be derived from the objectives for multimodal VAE [29], VRNN [28], and VAE for classification [30].

$$\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T} \lambda_1 \log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{<t}) + \lambda_2 \log p_\theta(y_t|z_{\leq t}, \mathbf{x}_{<t})\Big] - \sum_{t=1}^{T} \beta D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}), p_\theta(z_t)\big) \tag{5}$$

where λ_1, λ_2, β are the weights balancing the terms.

Algorithm 1 The proposed model

1: Initialize parameters of the generative model θ, recognition model ϕ, sequence length T.
2: Initialize optimizer parameters: $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\eta = 0.001$, $\epsilon = 10^{-10}$.
3: Initialize $W_0 \leftarrow \mathbf{1}$ and $x_1 \leftarrow \psi(X_1, W_0)$, where W_0 are the weights for the initial sampling, and the function ψ generates a sample x_1 from the environment X_1 after assigning weights W_0 (ref. Action selection in Section 2.3).
4: **while** true **do**
5:     **for** $\tau \leftarrow 1\ to\ T$ **do**
6:         $\hat{X}_{\tau+1}, \hat{y}_\tau \leftarrow Predict(x_{1:\tau})$
        Action execution (ref. Eq. 2–4)
7:         $x_{\tau+1} \leftarrow F(X_{\tau+1}, \hat{X}_{\tau+1})$
        Learning
8:         Update $\{\theta, \phi\}$ by maximizing Eq. 5.
9:     **end for**
10: **end while**

Algorithm 2 $Predict(x_{1:\tau})$

Recognition model (encoder)
1: $h_\tau^{enc} \leftarrow RNN_\phi^{enc}(x_\tau, h_{\tau-1}^{enc})$
2: $[\mu_\tau\,; \Sigma_\tau] \leftarrow \varphi^{enc}(h_{\tau-1}^{enc})$
3: $z_\tau | \mathbf{x}_\tau \sim \mathcal{N}(\mu_\tau, \Sigma_\tau)$

Generative model (decoder)
4: $h_\tau^{dec_1} \leftarrow RNN_\theta^{dec}(h_{\tau-1}^{dec_1}, z_\tau)$
5: $[\mu_{x,\tau}\,; \Sigma_{x,\tau}] \leftarrow \varphi^{dec}(h_\tau^{dec_1})$
6: $\hat{X}_{\tau+1} \leftarrow \mu_{x,\tau}$

Classification model (decoder)
7: $h_\tau^{dec_2} \leftarrow RNN_\theta^{dec}(h_{\tau-1}^{dec_2}, z_\tau)$
8: $\hat{y}_\tau \leftarrow softmax(h_{\tau-1}^{dec_2})$

3. Experimental Results

3.1. Datasets

We evaluate the model on two datasets: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Interactive Emotional Dyadic Motion Capture (IEMOCAP). We split each dataset into 80% train and 20% test.

RAVDESS [31] is an audiovisual dataset of emotional speech and songs. We consider the audio modality and the emotional speech, as in other works (e.g., [15, 17, 18, 20]). The considered data has 1440 audio files, vocalized by 24 professional actors (12 female, 12 male). The speech includes neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions.

The IEMOCAP is an acted, multimodal and multispeaker database, collected at SAIL lab at USC in five sessions. We consider the audio modality and the improvised speech data with four emotions (happy, sad, anger, neutral), as in [15]. The considered data has 2280 audio files, vocalized by 10 speakers (5 female, 5 male). We perform five-fold cross-validation by considering each session as one fold, as in [15, 19].

3.2. Experimental setup

Data preprocessing: From each audio file, we extract windows of 2 second duration, with 50% overlap. From each window, we extract 40 MFCCs using the Librosa library in Python. Thus, the input to our model at any instant t is a 40-dimensional MFCC vector. An audio file comprises of a sequence of T such vectors. Classification error is computed at the end of each audio file. We ignore windows of duration less than 2 seconds.

Training details: For RAVDESS, the recognition, generation and classification models consist of 512, 256, 512 hidden units respectively, and the latent variable dimension is 20. For IEMOCAP, the three models consist of 64 hidden units each, and the latent variable dimension is 10. These parameters are estimated experimentally. T is variable as the audio files are of different durations. We consider a minibatch size of 256. The parameters $\beta = 1$, $\lambda_1 = 1$, $\lambda_2 = 50$ are fixed. The model is learned end-to-end using backpropagation and RMSProp optimization with a learning rate of 0.001. These hyperparameters are estimated via cross-validation from the training set.

For RAVDESS, we use a dropout probability of 0.3 for recognition and generation hidden layers, and 0.1 for classification layer to prevent overfitting. For IEMOCAP, we use 0.8 for all three layers. The KL-divergence term in the objective function also acts as a regularizer [25] that prevents overfitting.

Ablation study: The utility of attention in our model is analyzed using an ablation study. We create a model by eliminating attention (Line 7 in Algorithm 1) from the proposed model. The VRNN is modified such that all locations in the observation are sampled with equal weight at any time. Thus, the model always observes the entire true MFCC vector. For a fair comparison, the number of layers and neurons in each layer for this non-attentional model is kept consistent with the original model.

Evaluation metrics: To evaluate recognition accuracy, we use the common metric, weighted accuracy, as computed in [18].[1] Efficiency of the model is evaluated in terms of the proportion of MFCC vector sampled for prediction.

Table 1: *Weighted classification accuracy (%).*

Data	Model (year)	Feature	WA
RAVDESS	Ours	MFCC	78.5
	Ours w/o attn.	MFCC	79.9
	[18] (2021)	MFCC	77.8
	[32] (2019)	Spectrogram	64.5
	[16] (2020)	Spectrogram	79.5[1]
	[17] (2021)	Spectrogram	80.0[1]
	[33] (2020)	Low-level features	72.0
IEMOCAP	Ours	MFCC	65.5
	Ours w/o attn.	MFCC	64.4
	[15] (2020)	IS09 + Mel spectrogram	67.7
	[5] (2018)	Spectrogram	70.4
	[19] (2018)	Low-level descriptors	68.8
	[21] (2017)	logMel filterbanks	62.1

3.3. Evaluation results

Evaluation for accuracy: We compare the classification accuracy of our model with recent works that use similar experimental setup as ours (ref. Table 1). Accuracy of our model is comparable to the state-of-the-art for RAVDESS dataset. Our model yields higher accuracy when compared to the state-of-the-art models using MFCC as features for RAVDESS. Accuracy of our model is comparable to some of those reported for IEMOCAP dataset. The number of parameters in our model is 22.6M

[1] Computation of unweighted accuracy in [16, 17] is similar to that of weighted accuracy in [18].

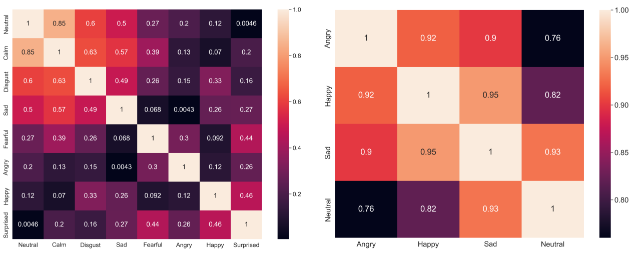
Table 2: *Percentage of MFCCs (total 40) deemed salient enough to be sampled by our model from the ground truth.*

Dataset	Neutral	Happy	Sad	Angry	Fearful	Disgust	Surprised	Calm	Average
RAVDESS	65.8	66.1	62.5	61.3	63.1	61.6	68.5	61.5	63.8
IEMOCAP	47.1	44.7	44.2	44.9	–	–	–	–	45.2

and 12.78M for RAVDESS and IEMOCAP respectively. The latter appears to be much less than the model in [5], which reported the state-of-the-art accuracy for IEMOCAP. Most works in this area did not report the number of model parameters, which makes fair comparison a challenge. Increasing our model size does not increase accuracy by much.

Generation error guides the attention of our model, which leads to efficiency by allowing selective sampling of the input observation. Hence, generation capability is a necessity in our model. The ablation study reveals that the classification accuracy of our model with and without attention are comparable, for both datasets (ref. Table 1). This shows that the generation and classification pathways do not interact adversely, and that the dynamic threshold used to compute attention weights (ref. Eq. 3) is a good choice to balance accuracy and efficiency.

Analysis of action selection: We visualize the similarity of attention weights between emotion classes in Fig. 2. For each emotion, the mean of weight vectors assigned to the MFCC vectors is computed. This mean weight vector for an emotion corresponds to the expected attention to the different MFCC features from the proposed model. The similarity of a pair of mean weight vectors is computed as the absolute of their normalized cosine similarity. For RAVDESS, 'neutral' is most similar to 'calm' and least to 'surprised'. For IEMOCAP, 'neutral' is most similar to 'sad' and least to 'angry'.



(a) RAVDESS dataset (b) IEMOCAP dataset

Figure 2: *These symmetric matrices show the similarity between the (expected) attention for each pair of emotion classes.*

Evaluation for efficiency: We compute the average (over all audios for each emotion) proportion of input observation (MFCC vector) sampled by our model at each time step (ref. Table 2). On average, for any interaction, our model samples 63.8% and 45.2% of the observation for RAVDESS and IEMOCAP respectively. The highest sparsity is for 'angry' emotion from RAVDESS and 'sad' emotion from IEMOCAP. The proposed model samples its observations efficiently without compromising accuracy.

4. Discussion

Attention. The attention mechanism in our model differs from most SER models from behavioral and algorithmic perspectives. Typically, end-to-end attention-based models for SER learn all parameters (including attention weights) by optimizing an objective function. In most of these models, attention is an internal mechanism that does not have a corresponding behavior. The attention parameters play a role similar to any other parameter in the model. In our model, attention is a parameterless mechanism that emerges due to prediction error, which drives action/behavior (ref. Eq. 2–4). This mechanism is interpretable as the model simply attends to its unexpected observations.

From an algorithmic perspective, SER models utilize attention weights at a higher feature level (e.g., [4–9, 11, 12, 17, 21]), or at multiple feature levels (e.g., [17, 19]). Our model utilizes attention at the input level only.

Many SER models compute attention weights from multiple time steps (e.g., [4, 6–8, 11, 12, 19, 21]). Such models need to process the input sequence till the final time, which introduces latency. Our model computes attention from the current time only. It infers by processing the input till the current time, which allows it to be efficient.

Speech representation. Even though the authors in [21] downplayed the role of speech representation, we believe the performance of our model can be improved by using a spectrogram instead of MFCCs. The spectrogram is less compact and contains more information than MFCC. Hence, using the spectrogram might increase the size and enhance the accuracy of our model. Most SER models with state-of-the-art accuracy use spectrogram.

ML component. VRNN-based models are known to perform well in computer vision applications (e.g., [34–36]). However, VRNNs have rarely been explored for SER. Our model using VRNN yields encouraging results. Convolution has been shown to be an effective operation for SER. Our model's accuracy might be improved by using a convolutional VRNN.

Generation and classification are done jointly in very few end-to-end ML models. The models reported in [29, 30] generate and classify handwritten numerals. Classification accuracy is not reported in [29]. In [24], a sentiment analysis model was proposed to generate and classify positive and negative reviews. The model yields lower classification error by jointly minimizing classification and generation losses than that by minimizing the former only. None of these models incorporate attention. For the problem of SER, the proposed model is the first to explore simultaneous generation and recognition.

5. Conclusions

We propose an attention-based predictive model for SER, where the key ML component is a VRNN. For efficiency, MFCC is used as the input speech representation. This model is the first to explore simultaneous generation and classification for SER. Generation error guides the attention of the model, leading to efficiency by selective sampling. The sampled observations are used for classification. Hence, the model performs recognition via generation. Our experiments using two benchmark datasets reveal that the model yields high recognition accuracy without compromising efficiency in terms of model size and sparsity of sampled observations. Using spectrogram as speech representation and a convolutional VRNN as the ML component might improve the accuracy but reduce the efficiency of this model.

6. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[2] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Commun.*, vol. 116, pp. 56–76, 2020.

[4] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *ICASSP*, 2017, pp. 2227–2231.

[5] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, “Attention based fully convolutional network for speech emotion recognition,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2018, pp. 1771–1775.

[6] M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, 2018.

[7] P.-W. Hsiao and C.-P. Chen, “Effective attention mechanism in dynamic models for speech emotion recognition,” in *ICASSP*, 2018, pp. 2526–2530.

[8] S. Yoon, S. Byun, S. Dey, and K. Jung, “Speech emotion recognition using multi-hop attention mechanism,” in *ICASSP*, 2019, pp. 2822–2826.

[9] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, “Speech emotion classification using attention-based LSTM,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, 2019.

[10] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, “Towards discriminative representation learning for speech emotion recognition.” in *IJCAI*, 2019, pp. 5060–5066.

[11] S. B. Alex, L. Mary, and B. P. Babu, “Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features,” *Circuits, Syst. Signal Process.*, vol. 39, no. 11, pp. 5681–5709, 2020.

[12] W.-C. Lin and C. Busso, “An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks,” in *Interspeech*, 2020, pp. 2322–2326.

[13] M. Seo and M. Kim, “Fusing visual attention CNN and bag of visual words for cross-corpus speech emotion recognition,” *Sensors*, vol. 20, no. 19, p. 5559, 2020.

[14] S. Mao, P. Ching, C.-C. J. Kuo, and T. Lee, “Advancing multiple instance learning with attention modeling for categorical speech emotion recognition,” *arXiv:2008.06667*, 2020.

[15] Y. Yu and Y.-J. Kim, “Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database,” *Electronics*, vol. 9, no. 5, p. 713, 2020.

[16] S. Kwon *et al.*, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, vol. 20, no. 1, p. 183, 2020.

[17] ——, “Att-Net: Enhanced emotion recognition system using lightweight self-attention module,” *Applied Soft Computing*, vol. 102, p. 107101, 2021.

[18] M. Xu, F. Zhang, and W. Zhang, “Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset,” *IEEE Access*, vol. 9, pp. 74 539–74 549, 2021.

[19] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, “Context-aware attention mechanism for speech emotion recognition,” in *IEEE Spoken Language Technology Workshop*, 2018, pp. 126–131.

[20] S. Kwon *et al.*, “CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network,” *Mathematics*, vol. 8, no. 12, p. 2133, 2020.

[21] M. Neumann and N. T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” *arXiv:1706.00612*, 2017.

[22] S. Latif, R. Rana, J. Qadir, and J. Epps, “Variational autoencoders for learning latent representations of speech emotion: A preliminary study,” *arXiv:1712.08708*, 2017.

[23] O. Maron and T. Lozano-Pérez, “A framework for multiple-instance learning,” *NIPS*, vol. 10, 1997.

[24] D. L. Marino, K. Amarasinghe, and M. Manic, “Simultaneous generation-classification using LSTM,” in *IEEE Symp. Ser. Comput. Intell.*, 2016, pp. 1–8.

[25] D. Kingma and M. Welling, “Auto-encoding variational Bayes,” *arXiv:1312.6114*, 2013.

[26] K. Friston, “The free-energy principle: A unified brain theory?” *Nat. Rev. Neurosci.*, vol. 11, no. 2, p. 127, 2010.

[27] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, “DRAW: A recurrent neural network for image generation,” *arXiv:1502.04623*, 2015.

[28] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, “A recurrent latent variable model for sequential data,” in *NIPS*, 2015, pp. 2980–2988.

[29] M. Wu and N. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” in *NIPS*, 2018, pp. 5575–5585.

[30] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” in *NIPS*, 2014, pp. 3581–3589.

[31] S. R. Livingstone and F. A. Russo, “The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[32] Y. Zeng, H. Mao, D. Peng, and Z. Yi, “Spectrogram based multi-task audio classification,” *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, 2019.

[33] H. Li, W. Ding, Z. Wu, and Z. Liu, “Learning fine-grained cross modality excitement for speech emotion recognition,” *arXiv:2010.12733*, 2020.

[34] M. Baruah and B. Banerjee, “A multimodal predictive agent model for human interaction generation,” in *CVPR Workshops*, 2020.

[35] ——, “The perception-action loop in a predictive agent,” in *CogSci*, 2020, pp. 1171–1177.

[36] M. Baruah, B. Banerjee, and A. K. Nagar, “An attention-based predictive agent for static and dynamic environments,” *IEEE Access*, vol. 10, pp. 17 310–17 317, 2022.